Aplicación de Técnicas Aprendizaje Automático para Estimar la Calidad de la Voz en Escala GRBAS

García Mario Alejandro[1], Rosset Ana Lorena[2], Moyano Miguel Alejandro[1], Ramírez Héctor Emilio[1], Melgralejo Samara Ofelia[1], Carrillo Florencia Noel[1]

[1]Universidad Tecnológica Nacional Facultad Regional Córdoba (UTN FRC)
[2]Universidad Nacional de Córdoba (UNC)

RESUMEN

La valoración de la calidad vocal mediante el análisis audio-perceptual es parte de la rutina clínica de evaluación de pacientes con trastornos de la voz. La debilidad de este método reside en la subjetividad y en la necesidad de que sea realizada por oyentes experimentados. Este proyecto tiene como objetivo particular la realización de una clasificación automática de la calidad vocal, valuada en la escala GRBAS, a través de características extraídas del análisis acústico de la señal y técnicas de aprendizaje automático.

Palabras clave: ***machine learning, deep learning, voice quality*, GRBAS**

CONTEXTO

Este trabajo de investigación se desarrolla en el marco del proyecto “Análisis acústico de la voz con técnicas de aprendizaje automático” Nacional, Facultad Regional Córdoba y cuenta con la colaboración del Departamento de Investigación Científica, Extensión y Capacitación "Raquel Maurette", Escuela de Fonoaudiología, Facultad de Ciencias Médicas, Universidad Nacional de Córdoba.

1. INTRODUCCIÓN

Se intenta reconocer, de forma automática, características del análisis acústico de la voz que permitan clasificar muestras de audio. El estudio se enfoca en la medición de la calidad vocal según la escala GRBAS. La clasificación se realizará aplicando principalmente modelos de deep learning, un subgrupo de técnicas del campo de aprendizaje automático (machine learning). Las grabaciones de la voz, la clasificación de los ejemplos y la validación de los resultados se realizarán por especialistas en análisis de la voz de la Escuela de Fonoaudiología de la Universidad Nacional de Córdoba. El análisis acústico se realizará en conjunto (especialistas vocales e integrantes de UTN) y el modelado y desarrollo de los clasificadores por los integrantes de UTN.

GRBAS: La escala GRBAS es un método de valoración perceptivo- auditivo de la voz. Surge de la necesidad de estandarizar la interrelacionar los aspectos auditivos y fisiológicos de la producción vocal. Está basada en estudios del año 1966 de la *Japan Society of Logopedics and Phoniatrics* [1] y posteriormente divulgada y descripta por Minoru Hirano en el año 1981 [2]. Consiste en la valoración de la fuente glótica a través de 5 parámetros que forman el acrónimo GRBAS:

G: (Grade) Grado general de disfonía R: (*Roughness*) Rugosidad, irregularidad de la onda glótica.

B: (*Breathiness*) Soplosidad, sensación de escape de aire en la voz.

A: (*Asteny*) Astenia, pérdida de potencia.

S: (*Strain*) Tensión, sensación de hiperfunción vocal.

Puede valorarse de dos maneras: a través de 4 grados, desde el 0 al 3 o mediante un valor en un rango continuo de 0 a 100. En ambas el 0 es ausencia de disfonía y el 3 o 100 implican disfonía severa. La escala fue mundialmente adoptada y validada en numerosos países [3], [4], [5], [6]. Actualmente se utiliza en la investigación y de manera rutinaria en los consultorios de los profesionales que hacen clínica vocal. Sirve como metodología simple y al alcance de la mano para valorar la evolución pre- post tratamiento. La debilidad de este método reside en la subjetividad de la valoración de la voz y en la necesidad de que sea realizada por oyentes experimentados en la escucha y la disociación de los parámetros [7], [8].

Análisis acústico: Existen otras formas de analizar la voz de manera más objetiva a través del análisis acústico. Éste consiste en la digitalización de la señal vocal y su análisis mediante gráficos como el Espectrograma, el espectro FFT (*Fast Fourier Transform*) o LPC (*Linear Predictive Coding*) y medidas numéricas de perturbación de la señal, como Jitter, Shimmer y HNR (*Harmonics to Noise Ratio*). El análisis acústico, a pesar de ser más objetivo, siempre necesita de la intervención del evaluador y esto es lo que también le impone una cierta subjetividad. Por este motivo surge la necesidad de la estandarización de todos los pasos en los que el sujeto evaluador va a intervenir: elección del material de habla que se graba, grabación, elección de los análisis a realizar (no se puede analizar todas las voces, más o menos disfónicas, con las mismas medidas) y finalmente, el análisis físico- acústico y fisiológico de los datos obtenidos. Para lograr una integración de la valoración subjetiva (GRBAS u otras escalas) con el análisis acústico, se han realizado numerosos trabajos de correlación [9], [10], algunos relacionados a la voz normal y otros a diferentes patologías. Por ejemplo, el trabajo de Nuñez Batalla, F. et al [11] es un referente y establece una relación entre el parámetro de Astenia del GRBAS y el Espectrograma de banda angosta.

Aprendizaje automático: El aprendizaje automático o *machine learning* es un campo de las ciencias de la computación que abarca el estudio y la construcción de algoritmos capaces de aprender y hacer predicciones. Estas predicciones se pueden tomar como una clasificación de los datos de entrada a partir del reconocimiento de patrones existentes en los mismos. Existen varios enfoques de *machine learning*. Estos difieren en el objetivo, tipo de entrenamiento, inspiración (por ejemplo matemática, estadística, biológica, etc.), eficiencia y complejidad entre otras características. Algunos de estos enfoques son redes neuronales artificiales, reglas de asociación, máquinas de vectores de soporte (*support vector machines*), árboles de decisión, redes bayesianas y análisis de *clusters*. *Deep Learning* es una rama del aprendizaje automático. Está compuesto por un grupo de algoritmos que intentan clasificar los datos en abstracciones de alto nivel mediante el uso de estructuras jerárquicas complejas. Algunas de las técnicas son *deep neural networks*, *convolutional neural networks* y *deep belief networks*.

Aprendizaje automático y análisis acustico: Los trabajos más importantes de los últimos años sobre *machine learning* y análisis acústico tienen como objetivo reconocer lo que se dice (*speech recognition*) y quién lo dice (*speaker recognition*). Los datos más frecuentemente utilizados como entrada a los modelos de *machine learning* son los Coeficientes Cepstrales en las Frecuencias de Mel (MFCC) o Coeficientes de Predicción Lineal Perceptual (PLPs) calculados directamente sobre la señal acústica y sobre sus primer y segunda derivada [12].

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Para el desarrollo de esta línea de investigación se toman tanto los últimos avances del campo de reconocimiento del habla, como la totalidad de los datos que brinda el análisis acústico clásico. Los valores devueltos por las técnicas de análisis acústico y los datos de originales (los valores muestreados que permanecen en los archivos de audio) son entradas potenciales para entrenar modelos de redes neuronales diseñados especialmente. El diseño implica la determinación del tipo de patrones que debe ser reconocido en cada una de las capas de la red, donde los primeros niveles reaccionan a estímulos de estructura simple y los más profundos lo hacen a las relaciones más complejas. El enfoque principal del proyecto es obtener en las primeras etapas de un modelo de *deep learning* valores equivalentes a los resultados del análisis acústico y buscar una correlación entre estos valores y los niveles de calidad vocal (GRBAS) en las últimas etapas del modelo.

Se programaron las siguientes tareas:

Grabación y entrevista:

Se realizará el protocolo de grabación de voz sugerido por Dejonckere *et al* [13] y se agregarán 2 aspectos extra:

1- una vocal /A/ sostenida en un tono e intensidad cómodos,

2- una vocal /A/ en una intensidad levemente ascendida respecto a la normal,

3- una frase simple estandarizada,

4- un fragmento de habla encadenada (serie automática) y

5- un glissando con la vocal /A/.

Lugar:

Departamento de Investigación Científica, Extensión y Capacitación "Raquel Maurette"

Entrevista:

Sobre cada persona se registrarán datos como edad, profesión/ocupación, género, antecedentes de problemas vocales, etc. Estos datos se colectarán en formularios predefinidos para luego cargarse en una base de datos que guarde la relación con los archivos de audio y los análisis acústicos posteriores.

Clasificación de las grabaciones

Los especialistas clasificarán según la escala GRBAS cada una de las grabaciones y se cargarán los datos en la base de datos.

Análisis acústico

Para cada grabación se realizará un análisis acústico (Espectrograma de banda angosta, Espectro FFT, Jitter y HNR) y se cargarán los resultados en la base de datos.

Software: Praat

Creación de modelos de aprendizaje automático (con datos de análisis acústico como entrada)

Se desarrollarán, entrenarán y evaluarán los modelos de aprendizaje en el lenguaje Python (principalmente con la librería scikit-learn).

Creación de modelos de deep learning (con audio y/o espectrograma como entrada)

Se desarrollarán, entrenarán y evaluarán los modelos de aprendizaje en el lenguaje Python (principalmete con el framework Keras sobre Theano).

3. RESULTADOS ESPERADOS

La estimación de la calidad vocal lograda de forma automática se debe comparar con la estimación de los expertos. Es importante notar que la clasificación puede diferir entre especialistas.

El nivel de acuerdo entre los especialistas de la voz se calculará con el coeficiente kappa de Cohen [72X], al igual que en el trabajo de De Bodt *et al* [73X].

El acuerdo entre la clasificación automática y los especialistas humanos también se calculará con el coeficiente kappa de Cohen, como Villa-Cañas *et al* en [74X].

Además, se calculará la desviación del modelo obtenido respecto a la media de las clasificaciones y se comparará con las desviaciones de los especialistas. Se espera obtener valores de variación significativamente menores.

4. FORMACIÓN DE RECURSOS HUMANOS

El equipo del proyecto está formado por un docente/investigador de la UTN FRC, dos docentes/investigadores de la UNC y cuatro alumnos de la carrera de grado de la UTN FRC.

Además de formación de los alumnos participantes, el conocimiento generado por el proyecto se incorporará a las cátedras de los docentes de la UTN y UNC.

5. REFERENCIAS

[1] Isshiki, N., Yanagihara, N., & Morimoto, M. (1966). *Approach to the objective diagnosis of hoarseness*. Folia Phoniatrica et Logopaedica, 18(6), 393- 400.

[2] Hirano, M. (1981). *Clinical examination of voice* (Vol. 5). Springer.

[3] Yun, Y. S., Lee, E. K., Baek, C. H., & Son, Y. I. (2005). *The correlation of GRBAS scales and laryngeal stroboscopic findings for the assessment of voice therapy outcome in the patients with vocal nodules*. Korean Journal of Otolaryngology- Head and Neck Surgery, 48(12), 1501- 1505.

[4] Hui, H., Weijia, K., & Shusheng, G. (2007). *The Validation of Acoustic Analysis and Subjective Judgment Scales of Several Voice Disorders* [J]. Journal of Audiology and Speech Pathology, 3, 010.

[5] Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (2007). *Reliability of clinician- based (GRBAS and CAPE- V) and patient- based (V- RQOL and IPVI) documentation of voice disorders*. Journal of Voice, 21(5), 576- 590.

[6] Jesus, L. M., Barney, A., Couto, P. S., Vilarinho, H., & Correia, A. (2009, December). *Voice quality evaluation using cape- v and GRBAS in european Portuguese*. In MAVEBA (pp. 61- 64).

[7] Kreiman, J., & Gerratt, B. R. (2010). *Perceptual assessment of voice quality: past, present, and future*. SIG 3 Perspectives on Voice and Voice Disorders, 20(2), 62- 67.

[8] Núñez- Batalla, F., Díaz- Molina, J. P., García- López, I., Moreno- Méndez, A., Costales- Marcos, M., Moreno- Galindo, C., & Martínez- Camblor, P. (2012). El espectrograma de banda estrecha como ayuda para el aprendizaje del método GRABS de análisis perceptual de la disfonía. Acta Otorrinolaringológica Española, 63(3), 173-179.

[9] Freitas, S. V., Pestana, P. M., Almeida, V., & Ferreira, A. (2015). *Integrating Voice Evaluation: Correlation Between Acoustic and Audio- Perceptual Measures*. Journal of Voice, 29(3), 390- e1.

[10] ELISEI, N. G. (2013, May). Percepción auditiva de voces patológicas. In XIV Reunión Nacional y III Encuentro Internacional De La Asociación Argentina de Ciencias del Comportamiento.

[11] Nuñez Batalla, F., Corte Santos, P., Señaris Gonzalez, B., Rodriguez Prado, N., Suárez Nieto, C. (2004) Evaluación espectral de la hipofunción vocal. Acta Otorrinolaringol. Esp. 55:327- 333.

[12] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*. Signal Processing Magazine, IEEE, 29(6), 82-97.

[13] Dejonckere, P. H., Bradley, P., Clemente, P., Cornut, G., Crevier- Buchman, L., Friedrich, G., ... & Woisard, V. (2001). *A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques*. European Archives of Oto- rhinolaryngology, 258(2), 77- 82.

[14] Cohen J. *A coefficient of agreement for nominal scales*. Educ Psych Measurement 1960;20:37—46.

[15] De Bodt, M. S., Wuyts, F. L., Van de Heyning, P. H., & Croux, C. (1997). *Test-retest study of the GRBAS scale: influence of experience and rofessional background on perceptual rating of voice quality*. Journal of Voice, 11(1), 74-80.

[16] Villa-Cañas, T., Orozco-Arroyave, J. R., Arias-Londono, J. D., Vargas-Bonilla, J. F., & Godino-Llorente, J. I. (2013, September). *Automatic assessment of voice signals according to the GRBAS scale using modulation spectra, Mel frequency Cepstral Coefficients and Noise Parameters*. In Image, Signal Processing, and Artificial Vision (STSIVA), 2013 XVIII Symposium of (pp. 1-5). IEEE.